

An Interesting Property of LPCs for Sonorant Vs Fricative Discrimination

T. V. Ananthapadmanabha¹, A. G. Ramakrishnan^{2*}, SMIEEE, Pradeep Balachandran³, MIEEE

Abstract—Linear prediction (LP) technique estimates an optimum all-pole filter of a given order for a frame of speech signal. The coefficients of the all-pole filter, $1/A(z)$ are referred to as LP coefficients (LPCs). The gain of the inverse of the all-pole filter, $A(z)$ at $z = 1$, i.e., at frequency = 0, $A(1)$ corresponds to the sum of LPCs, which has the property of being lower (higher) than a threshold for the sonorants (fricatives). When the inverse-tan of $A(1)$, denoted as $T(1)$, is used as a feature and tested on the sonorant and fricative frames of the entire TIMIT database, an accuracy of 99.07% is obtained. Hence, we refer to $T(1)$ as sonorant-fricative discrimination index (SFDI). This property has also been tested for its robustness for additive white noise and on the telephone quality speech of the NTIMIT database. These results are comparable to, or in some respects, better than the state-of-the-art methods proposed for a similar task. Such a property may be used for segmenting a speech signal or for non-uniform frame-rate analysis.

Keywords—Linear prediction, phonetic classes, manner classes, V/U classification, segmentation, SFDI

I. INTRODUCTION

Integration of knowledge of phonetic classes into a statistical based ASR system [1]–[3] is known to supplement its performance. This paper is concerned with the discriminability between two important phonetic classes, viz, sonorants and fricatives, from a continuous speech signal. The class of ‘sonorants’ comprises vowels and voiced consonants excluding voiced stops (b, d, g) - all voiced with the only exception being ‘hh’ which is unvoiced. The class of fricatives comprises unvoiced phones ‘s’, ‘sh’, ‘f’, ‘ch’ and mixed voiced-unvoiced phones ‘z’, ‘zh’, ‘jh’, ‘th’, ‘dh’ and ‘v’.

In the literature, this problem has been studied under the context of manner classification and landmark detection [4]–[6] and also in the context of extraction of distinctive features [7]–[9]. The problem of identifying sonorants Vs unvoiced fricatives may also be looked upon as a V/U classification problem, which has been extensively studied in the literature and we refer to some of them here [10]–[17].

Several methods have been proposed for the identification of broad phonetic classes and/or their onsets from a speech signal. Liu [5] has used the change of energy between two frames spaced 50 ms apart, over six sub-band signals, for detecting the onsets or landmarks of four broadly defined classes. Salomon et al [6] have used a set of twelve temporal parameters to achieve manner classification. A team of

researchers have used landmark based approach [7] for feature extraction and experimented with different classifiers, such as SVMs, for identifying the distinctive features, which in turn may be used for manner classification. King and Taylor have used mel-frequency cepstral coefficients (MFCCs) and their derivatives (a 39-dimensional feature vector) to train a neural network to identify multi-dimensional distinctive features comprising broad manner classes [8]. Juneja and Wilson combined MFCCs with certain temporal features and used an SVM classifier for the manner classification task [9].

Most of the methods on V/U classification use the following temporal features [11]: (i) the relative energy of a frame (which is typically low for unvoiced frames), (ii) the ratio of energies in the lowpass to highpass region (which is typically high for voiced segments), (iii) the number of zero-crossings per unit interval (which is typically high for the unvoiced segments), (iv) the value of normalized autocorrelation at one sample lag, which indirectly relates to the first reflection coefficient in linear prediction (LP) analysis captures the gross spectral slope (typically lowpass for voiced and highpass for unvoiced), (v) periodicity detection (voiced sounds are periodic) and (vi) pitch prediction gain. Deng and O’Shaughnessy [16] have used an unsupervised algorithm for V/U classification and tested the performance on the NTIMIT database. Other features have also been considered. Alexandru Caruntu et al [13] have used zero-crossing density, Teager energy and entropy measures. Dhananjaya and Yegnanarayana [17] have used glottal activity detection, which in turn requires epoch extraction. Molla et al. [18] have modeled the speech signal as a composite signal of intrinsic mode functions and extracted the trend of these functions. The trends are compared with thresholds obtained on a training data for V/U classification.

In this paper, we demonstrate that the sum of linear prediction coefficients (LPCs), a scalar measure, is useful in discriminating a sonorant class from a fricative class in a speech signal. LP is a very successful speech analysis technique [19]–[21]. According to the frequency domain interpretation [21], LP technique estimates an optimum all-pole digital filter, $1/A(z)$, that best approximates the short-time spectrum of a frame of speech signal. The reciprocal all-zero filter, $A(z)$, called the digital inverse filter is given by

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + \dots + a_M z^{-M} \quad (1)$$

where a_1, a_2, \dots, a_M are the LPCs and M is the number of LPCs, which is a variable that can be set during the estimation of LPCs. Here z^{-1} is the unit delay operator given by $z^{-1} = \exp^{-j2\pi fT}$, where T is the sampling interval. Hence, the gain of the filter $A(z)$ at $z = 1$ (or frequency = 0), i.e., $A(1)$ is given

¹Voice and Speech Systems, Malleswaram, Bangalore, India 560003 (e-mail: tva.blr@gmail.com).

^{2*}Department of Electrical Engineering, Indian Institute of Science, Bangalore, India 560012 (e-mail: ramkiag@ee.iisc.ernet.in).

³Department of Electrical Engineering, Indian Institute of Science, Bangalore, India 560012 (e-mail: pb@mile.ee.iisc.ernet.in).

by

$$A(z=1) = 1 + a_1 + a_2 + a_3 + \dots + a_M \quad (2)$$

which corresponds to the sum of the LPCs.

In this paper, we report an interesting application of $A(1)$ for segmentation of a speech signal into broad phonetic classes and test its effectiveness using the entire TIMIT database.

II. $A(1)$ CONTOUR AND ITS CHARACTERISTIC

Speech signal is divided into frames of 20 ms with two successive frames spaced by 5 ms. A Hanning window is applied on a frame of speech signal after removing the mean value for the frame. The autocorrelation method of LP technique [20] is used as it assures the filter to be stable. The computed LPCs depend on the spectral shape, but are independent of the signal level. The number of LPCs is chosen as $(F_s + 2)$, where F_s is the sampling frequency in kHz. LPCs are computed on the preemphasized and windowed speech signal. After obtaining the LPCs, the computed $A(1)$ is assigned to the entire mid 5 ms of the speech frame. Thus a contour of $A(1)$ appears like a staircase waveform. It has been found that the typical value of $A(1)$ is greater than 13 for fricatives and less than 0.4 for sonorants.

The rationale for selecting $A(1)$ as a feature arises as follows. Assuming the speech signal as the output of a quasi-stationary time-invariant filter (vocal tract) excited by an appropriate excitation, the intensity of a frame of speech signal is determined both by the strength of excitation (source intensity) and the filter gain. Given the speech signal, the source intensity can be obtained by an inverse filtering operation. It has been observed that the source intensity is higher for fricatives than for sonorants, which is exactly the opposite of what is observed for speech intensity. This may arise since there is a greater amount of acoustic loss for fricatives than for the resonant sounds of vowels. The filter gain itself is determined by two factors, namely, the frequency response and the filter gain at $f = 0$. Conventionally, a standard method of representing a filter is to set the filter gain at $f = 0$ to be unity. Hence we deduce that the filter gain at $f = 0$ must be influencing the source intensity. This led us to investigate $A(1)$ as an acoustic feature for distinguishing sonorants from fricatives.

An illustrative example: The utterance (sa2.wav) ‘Don’t ask me to carry an oily rag like that’ from the TIMIT database is analyzed. The speech wave and the computed $A(1)$ values are shown in Fig. 1a for a part of the utterance. The hand labeled boundaries and the phone labels are also shown in the figure. Here $A(1)$ is forced to zero for the silence frames, determined by a threshold on the energy. $A(1)$ rises sharply at the onset of the fricative ‘s’, reaches the maximum value of 28 and falls sharply at the end of the fricative segment. In Fig. 1a, we note that for sonorants, the maximum value of $A(1)$ is low (<1.0). We make use of this property of $A(1)$. In this paper, we study the characteristic of $A(1)$ for the 2-class problem of sonorants Vs fricatives. The maximum value of $A(1)$ observed for fricatives is 119 and the minimum is 0.058 for sonorants. The large value of $A(1)$ for fricatives dominates over the sonorants. The variation in $A(1)$ within

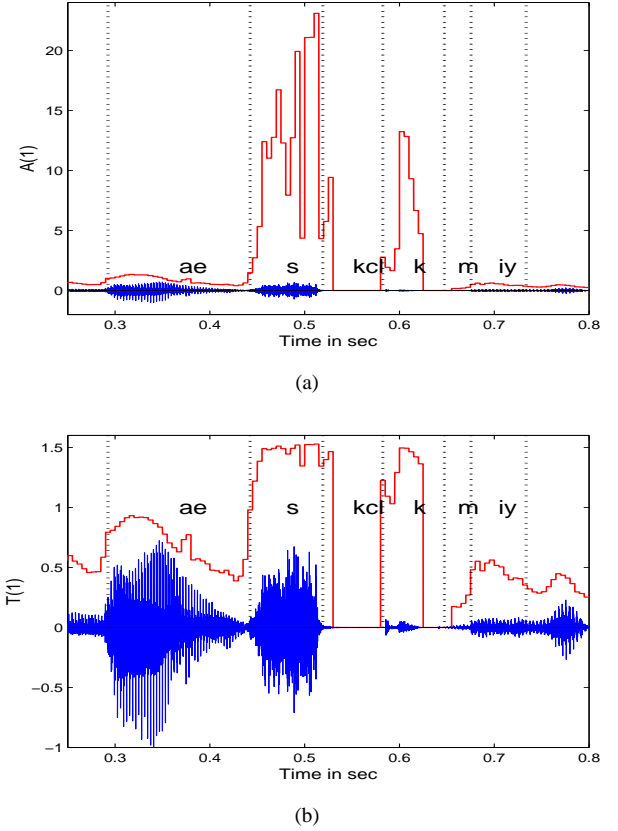


Fig. 1: Segment of a speech signal comprising the utterance ‘ask me’ (sonorant-fricative-stop-sonorant) and the plots of frame-wise values of (a) $A(1)$; (b) $T(1)$ in radians.

a fricative segment is not of interest as long as the value is above a threshold. Hence, we prefer to compress the range of $A(1)$ using

$$T(1) = \tan^{-1}[A(1)] \quad (3)$$

Such a compression of $A(1)$ also helps in graphic visualization in comparison with a normalized signal waveform plot. The upper bound for $T(1)$ is $\pi/2$. During our investigation, we did not come across a single instance, where $A(1)$ is negative. Fig. 1b shows the plot of $T(1)$, which compresses the range of $A(1)$ and also swamps out the variations when $A(1)$ is large. Henceforth, $T(1)$ is termed as sonorant-fricative discriminant index (SFDI).

For the stop segment (marked as ‘k’ in Fig. 1b), $T(1)$ reaches a maximum value close to 1.4 for a part of the segment. Stop bursts usually follow a silence or a low level voicing, which may be utilized for their detection [22]. Other phones also exhibit mixed characteristics for $T(1)$. However, the detection of stop bursts and other phones is not a topic for this paper.

III. EXPERIMENTS AND RESULTS

A. Histogram of SFDI for sonorants and fricatives

The hand labeled TIMIT database [23] is used for validating the proposed concept. The labeled boundaries in the TIMIT database are used to identify the sonorant and fricative segments. SFDI is computed frame-wise over these segments. For

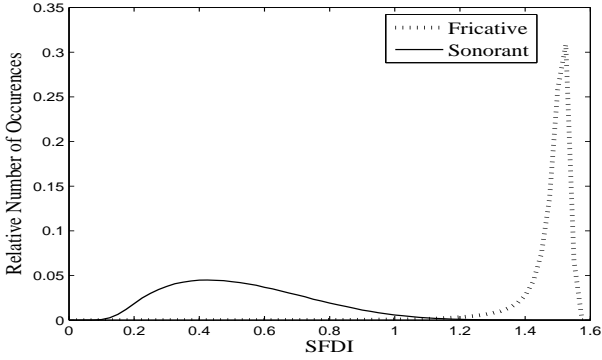


Fig. 2: Normalized histogram of the value of SFDI within sonorant (solid) and fricative (dashed) segments from the entire TIMIT database.

the purpose of computing histograms, ‘hh’ is excluded from sonorants since it behaves like an unvoiced sound. Phones ‘th’, ‘dh’ and ‘v’ are excluded from the fricative class, since these phones behave anomalously. The entire TIMIT database is analyzed. The value of SFDI is computed for all the frames within each segment of the two classes, sonorants and fricatives. The histograms of SFDI for the two classes, normalized by the number of occurrences of each class, are shown in Fig.2 for a bin size of 0.025. The two histograms show a clear separation with a very small overlap. The number of sonorants falls sharply for SFDI values above 1.1, whereas SFDI has values larger than 1.1 for the fricatives.

B. Arriving at the threshold and the frame-wise performance

Assume that a threshold based logic is used and whenever SFDI is less than a threshold T , the frame is assigned to the sonorant class; else, to the fricative class. If SFDI for a known frame of a sonorant (fricative) is lesser (greater) than the threshold T , then that frame is considered to be correctly classified. The ratio of the number of correctly classified frames to the total number of frames studied gives the accuracy. The frame-wise accuracy is computed for the two classes on a development set of the TIMIT database, consisting of 1140 files. As the threshold is increased, the error rate falls sharply for the sonorants since lesser number of sonorant frames have a higher value of SFDI. On the other hand, as the threshold is increased, the area under the normalized histogram below the threshold increases thereby increasing the error rate for the fricatives. Figure 3 shows the frame-wise error rate Vs threshold. The cross-over point of error rates occurs at a threshold of 1.1 and the corresponding error rate is about 0.80%. A similar method is used for arriving at the threshold for the entire TIMIT database consisting of 6300 files. The threshold is found to be 1.1, which is the same as that obtained for the development set. Frame-wise error rate with this threshold is 0.93%, which is about the same as that obtained for the development set. Some of the errors occur at the frame boundaries, where the hand labeled boundaries may not have been accurately marked.

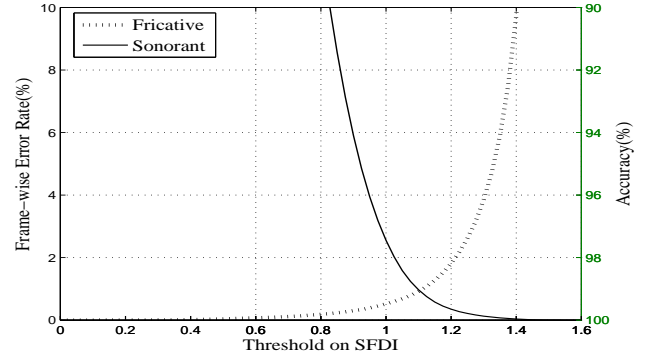


Fig. 3: Frame-wise error rate on the development set of TIMIT as a function of threshold on SFDI for sonorants (solid) and fricatives (dashed).

TABLE I: Effect of noise on the performance of SFDI on the development set and the entire database.

**Orgnl: Original data set of TIMIT and NTIMIT, supposed to have SNR values of 39.5 dB and 26.5 dB, respectively.

Noise Level	TIMIT			NTIMIT		
	T	Accuracy (%)		T	Accuracy (%)	
		Dev. Set	Full Set		Dev. Set	Full Set
Orgnl.**	1.10	99.20	99.07	0.62	89.04	88.63
20 dB	1.30	98.52	98.38	0.99	87.94	87.35
15 dB	1.36	98.21	98.04	1.12	87.10	86.59
10 dB	1.41	97.65	97.48	1.22	85.97	85.54
5 dB	1.43	95.78	95.65	1.29	84.05	83.91
0 dB	1.44	91.53	91.35	1.33	80.83	80.44

C. Extended experiments

It has been reported that the results obtained with a development set are comparable to that obtained with the entire database [6], [16], [22]. Here, we present the results for both the development set as well as for the full set of TIMIT.

1) *Robustness*: In order to study the robustness of the measure SFDI, white Gaussian noise is added to the speech samples with an appropriate scale factor to achieve the desired global SNR. SNRs of 20 down to 0 dB, in steps of 5 dB, are considered in the experiment. The value of the threshold T at the cross-over point is measured. The value of T is about the same for both the development set and the entire database. The corresponding frame-wise accuracy is shown in Table I.

As the SNR decreases, the threshold T increases. Assuming the white noise and speech signal to be uncorrelated, the spectral level shifts uniformly across all frequencies and hence SFDI also increases. The frame-wise accuracy even at 0 dB SNR is 91.3%, which is respectable.

2) *Experiments on the NTIMIT database*: The NTIMIT database is the telephone quality counterpart of the TIMIT database with a reduction in the bandwidth (300-3400 Hz) as well as a deterioration of SNR (39.5 to 26.8 dB) [24]. Once again, the value of T is noted to be the same for the development set and for the entire database. The frame-wise accuracy for the original speech is around 89%. The results for different SNRs are shown in Table I. The performance degrades with noise and the frame-wise accuracy is about

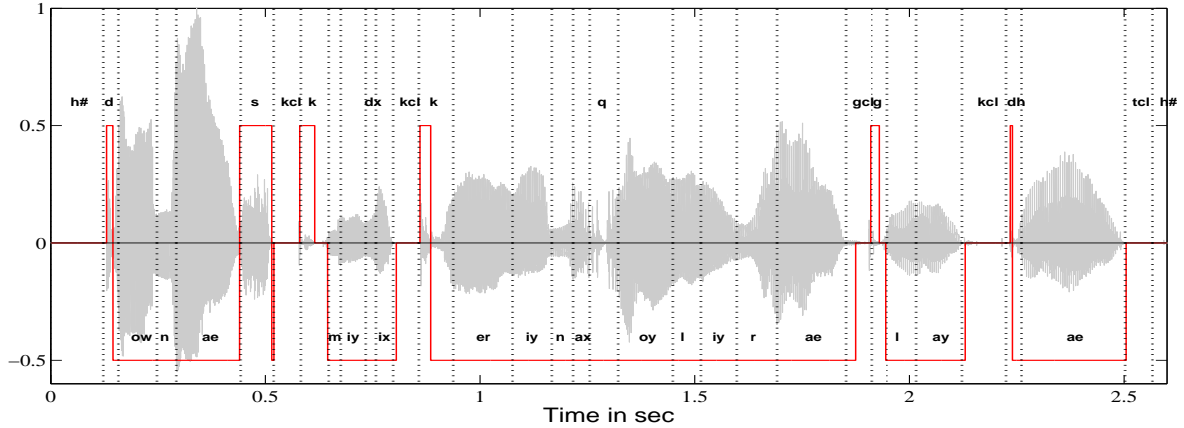


Fig. 4: Two-class segmentation of the utterance ‘Don’t ask me to carry an oily rag like that’ using SFDI.

80.4% at 0 dB SNR

3) *SFDI-based 2-class segmentation*: Using SFDI, speech signal is broadly segmented into two classes. The frames with $SFDI > T$ are assigned to class-1 and others to class-2. A rectangular function is defined with values of 0.5 and -0.5 for the two classes. If the energy in a frame is lower than 0.0004 times the maximum frame energy, then the value of the rectangular function is forced to zero for that frame to signify a silence segment. The resultant three-level rectangular function is shown in Fig.4 along with the speech waveform for an utterance from TIMIT. The hand-labeled boundaries and corresponding labels are shown. Identified class-1 mostly comprises fricatives and class-2 mostly comprises sonorants and the detected boundaries lie close to the hand labeled boundaries between the two classes. It may be noted that for stops, the segments are very short in duration and are preceded by a silence segment (closure duration). Since the likely interval of the bursts can be deduced, the plosion index used for detecting a burst [22] need to be computed only over such an interval instead of at every sample. This illustrates the potential use of SFDI for segmentation. Also, such a segmentation may be used for non-uniform frame rate analysis. A formal validation of segmentation, primarily using SFDI as a feature, is beyond the scope of the present paper.

D. Comparison with previous work

Although a strict comparison with previous studies is not possible since the size of the database used and the tasks addressed are different, we make some broad observations for comparative purposes and these must not be construed as a criticism of the earlier results.

For the manner classification task, the reported accuracy is in the range of 70-80% [6], [9]. Frame-wise accuracy for the distinctive feature voiced/unvoiced is reported to be about 93% and for vowel/fricative distinction, about 87% for the TIMIT database and original speech [8]. The accuracy for [sonorant] or [fricative] detection is on an average about 95% for TIMIT, using 42 acoustic parameters and SVM classifier [7]. The highest reported accuracies for V/U classification for

the TIMIT database are about 94.4% for original speech and 92.7, 91.8, 89.7 and 86.4% for 20, 10, 5 and 0 dB SNR, respectively for additive white noise [17]. For the NTIMIT database, the reported classification accuracy is about 80% based only on the algorithm, i.e., without correcting for the transcription errors [16]. In comparison, the accuracy obtained in this study for a broad 2-class separation using a single scalar measure of SFDI is about 99.07% for the original, 91.35% at 0 dB SNR for the TIMIT database and 88.6% for the original and 80.4% at 0 dB SNR for the NTIMIT database. It has been observed that the results obtained for a development set and the entire database are about the same.

IV. CONCLUSION

This study has demonstrated that a simple scalar measure SFDI, viz., inverse tan of sum of LPCs, along with a threshold based logic, may be effectively used to distinguish the sonorants from the fricatives. The experiments show that the discrimination given by SFDI is high even at 0 dB SNR. The results obtained in this study are comparable, or in some respects, better than the state-of-the-art methods.

Future research would be to utilize this property of SFDI, along with additional features, for automatic segmentation of a speech signal into different phonetic classes. It would be of interest to understand the reason for the observed property of SFDI by studying the relationship between the spectral envelope and the sum of LPCs.

REFERENCES

- [1] R. P. Lippmann, “Speech recognition by machines and humans,” *Speech communication*, vol. 22, no. 1, pp. 1–15, 1997.
- [2] W. J. Barry and W. A. van Dommelen, *The integration of phonetic knowledge in speech technology*. Springer, 2005, vol. 25.
- [3] P. Niyogi and P. Ramesh, “The voicing feature for stop consonants: recognition experiments with continuously spoken alphabets,” *Speech Communication*, vol. 41, no. 2, pp. 349–367, 2003.
- [4] K. N. Stevens, “Toward a model for lexical access based on acoustic landmarks and distinctive features,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1872–1891, 2002.
- [5] S. A. Liu, “Landmark detection for distinctive feature-based speech recognition,” *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3417–3430, 1996.

- [6] A. Salomon, C. Y. Espy-Wilson, and O. Deshmukh, "Detection of speech landmarks: Use of temporal information," *The Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1296–1305, 2004.
- [7] M. Hasegawa-J, J. Baker, S. Greenberg, K. Kirchhoff, J. Muller, K. Sonmez, S. Borys, K. Chen, A. Juneja, K. Livescu, S. Mohan, E. Coogan, and T. Wang, "Landmark-based speech recognition," *Report of the 2004 Johns Hopkins summer workshop*, 2005.
- [8] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 333–353, 2000.
- [9] A. Juneja and C. Espy-Wilson, "Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning," *Proc. IEEE Int. Conf. Neural Information Processing*, pp. 726–730, 2002.
- [10] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 3, pp. 201–212, 1976.
- [11] J. P. Campbell and T. E. Tremain, "Voiced/unvoiced classification of speech with applications to the U.S. government LPC-10E algorithm," *ICASSP, Tokyo*, vol. 11, pp. 473–476, April 1986.
- [12] A. P. Lobo and P. C. Loizou, "Voiced/unvoiced speech discrimination in noise using Gabor atomic decomposition," *Proc. Int. Conf. Acoustics Speech and Signal Processing, Hong Kong*, no. 1, pp. 820–823, Apr. 2003.
- [13] A. Caruntu, A. Nica, G. Todorean, E. Puschita, and O. Buza, "An improved method for automatic classification of speech," *IEEE International Conference on Automation, Quality and Testing Robotics*, vol. 1, pp. 448–451, 2006.
- [14] C. Shahnaz, W. P. Zhu, and M. O. Ahmad, "A multifeature voiced/nonvoiced decision algorithm for noisy speech," *Proc. Int. Symp. Circuits and Systems, Kos, Greece*, pp. 2525–2528, May 2006.
- [15] D. Arifianto, "Dual parameters for voiced-unvoiced speech signal de-termination," *Proc. Int. Conf. Acoustics Speech and Signal Processing, Honolulu*, no. IV, pp. 749–752, April 2007.
- [16] H. Deng and D. O'Shaughnessy, "Voiced-unvoiced-silence speech sound classification based on unsupervised learning," *IEEE International Conference on Multimedia and Expo*, pp. 176–179, 2007.
- [17] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Sig. Proc. Letters*, vol. 17, no. 3, pp. 273–276, March 2010.
- [18] M. K. I. Molla, K. Hirose, S. K. Roy, and S. Ahmad, "Adaptive thresholding approach for robust voiced/unvoiced classification," *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2409–2412, 2011.
- [19] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am*, no. 50, pp. 637–655, 1971.
- [20] J. Markel, "Digital inverse filtering—a new tool for formant trajectory estimation," *Audio and Electroacoustics, IEEE Transactions on*, vol. 20, no. 2, pp. 129–137, 1972.
- [21] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.
- [22] T. V. Ananthapadmanabha, A. P. Prathosh, and A. G. Ramakrishnan, "Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 460–471, 2014.
- [23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Acoustic-phonetic continuous speech corpus," *DARPA TIMIT, U.S. Department of Commerce, Washington, DC*, no. 4930, 1993.
- [24] Jankowski, Charles, Kalyanswamy, Ashok, Basson, Sara, Spitz, and Judith, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing (ICASSP), Albuquerque*, vol. 1, pp. 109–112, April 1990.